CLUSTER MERGING BASED ON WEIGHTED MAHALANOBIS DISTANCE WITH APPLICATION IN DIGITAL MAMMOGRAPHY

Khaled Younis, Mohammed Karim*, Russell Hardie*, John Loomis*, Steven Rogers**, Martin DeSimio****

* The University of Dayton, Dayton, OH 45469
** Battelle Memorial Institute, Columbus, OH 43201
*** Qualia Computing, Incorporated, Dayton, OH 45433

ABSTRACT

A new clustering algorithm that uses a weighted Mahalanobis distance as a distance metric to perform partitional clustering is proposed. The covariance matrices of the generated clusters are used to determine cluster similarity and closeness so that clusters which are similar in shape and close in Mahalanobis distance can be merged together serving the ultimate goal of automatically determining the optimal number of classes present in the data. Properties of the new algorithm are presented by examining the clustering quality for codebooks designed with the proposed method and another common method that uses Euclidean distance. The new algorithm provides better results than the competing method on a variety of data sets. Application of this algorithm to the problem of detecting suspicious regions in a mammogram is discussed.

1. INTRODUCTION

Clustering is a very important tool in pattern recognition for identifying structure in data. Clustering usually implies partitioning of a collection of objects (tanks, handwritten digits, cancerous areas in a mammogram) into c disjoint subsets. That is, to partition a set $\mathcal{H}$ of n samples $X = \{x_1, x_2, .., x_n\} \subset \mathcal{R}^d$ into subsets $\mathcal{H}_1, .., \mathcal{H}_c$. Each subset is to represent a cluster, with objects in the same cluster being somehow "more similar" than samples in different clusters. In other words, objects in a cluster should have common properties which distinguish them from the members of the other clusters. Each subset $\mathcal{H}_i$ is represented by a codeword v_i, where v_i is the centroid of the samples in $\mathcal{H}_i$.

A well known algorithm for the design of a locally optimal codebook with iterative codebook improvement is the generalized Lloyd algorithm (GLA) [1]. The two steps in each iteration of this algorithm are:

Step 1: Given a codebook $C_m = \{v_i \; ; \; i = 1,\ldots,k\}$ obtained from the m^{th} iteration, assign each data point to the *closest* codeword.

Step 2: Obtain the codebook C_{m+1} by computing the centroid of each cluster based on the partitioning of Step 1.

The above algorithm is usually terminated when the codewords stop moving or the difference between their locations in consecutive iterations is below a threshold.

The closest codeword is typically found with a distance metric. A general form of the distance between vectors x and a codeword v_i is

$$D = \|x - v_i\|_A^2 = (x - v_i)' A^{-1} (x - v_i) \qquad (1)$$

where A is any positive definite d x d matrix. The Euclidean distance, $D = \|x - v_i\|_I$, is a commonly used distance metric in practice, where I is the identity matrix.

The choice of an optimality criterion is a very important issue in the design of a clustering algorithm. Especially in higher dimensions, one cannot visually determine how good the resulting clusters are. One approach is to check with a criterion function. The criteria, or performance indices, are often specified as a function of the memberships whose minima or maxima define "good" clustering. The algorithm then becomes a numerical procedure for finding memberships which optimize the objective function.

The simplest and most widely used criterion function for clustering is the sum of squared error criterion. Bezdek [2] generalized a criterion function to account for the fuzzy membership values. The generalized mean squared error is defined by:

$$J_{GMSE} = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n_i} (u_{ij})^m \|x_j - v_i\|_A^2 \qquad (2)$$

0-7803-4449-9/98/$10.00 1998 IEEE

where u_{ij} is the membership of the i^{th} pattern to the j^{th} cluster, m is a weighting exponent strictly greater than one, and A is any positive definite matrix [2].

Choosing the Euclidean distance in clustering implies an isotropic feature space weighting [3]. This isotropic assumption tends to form hyperspherical clusters. Hence, clustering using the Euclidean distance may split large or elongated clusters. It is not uncommon for data to fall naturally into hyperellipsoids in the feature space rather than in hyperspheres.

An alternative distance metric that takes into consideration the distribution of the data is the Mahalanobis distance (MD)[1]. The MD between any input sample x and a codeword v_i is computed by evaluating (1) where A is the sample covariance matrix of the samples in $\mathcal{H}_i$. In our algorithm we propose the use of an individual covariance matrix A_i for each cluster and update A_i based on the partitioning after each iteration. The next section details the new approach.

2. WEIGHTED MAHALANOBIS DISTANCE (WMD) CLUSTERING

The idea behind the proposed method is to make each cluster attract those data points that enhance its own shape as implied by the covariance matrix of the samples within that cluster. In this algorithm, we modify the GLA described in the previous section such that in each iteration we assign a pattern x to the cluster that yields the minimum *weighted* Mahalanobis distance, $D_i = W_i * ||x - v_i||^2_{A_i}$. Where W_i is the cluster weight. In the second step, we update v_i and A_i by computing the mean and the covariance matrix of the data points of each cluster based on the partitioning of the first step. The algorithm is terminated when the codewords stop moving.

The introduction of the weight W is due to the fact that the use of Mahalanobis distance alone in clustering sometimes causes a large cluster to attract members of neighboring clusters. This leads to unusually large and unusually small clusters [4][5].

Looking at the J_{GMSE} criterion function again and allowing A_i to be variable

$$J_{GMSE} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}(u_{ij})^m(x-m)^T A_i^{-1}(x-m) \quad (3)$$

it is clear that there is a need to restrict A_i somehow in order to obtain a nontrivial solution. Otherwise, the minimum of J_{GMSE} would be given by A_i^{-1}=0, which corresponds to a huge cluster with infinite variation in all directions.

One way to solve this problem is to force the determinant of all clusters to have a unity value. Therefore, we evaluate W_i as

$$W_i = \frac{1}{|A_i|^{\frac{1}{d}}} \quad (4)$$

where $|A_i|$ is the determinant of A_i.

This choice of the constraint has the effect of normalizing the volume enclosed by the equi-Mahalanobis distance hyperellipsoid to a constant volume for all clusters while maintaining the distinct shape of each cluster.

To get an initial codebook we used the Karhunen-Loève transformation to place the initial codewords along the principal component axes of the data's covariance matrix. For the first iteration, we use the global covariance matrix as A_i for all the codewords. If the number of data points in any cluster is less than the dimensionality of the data, then A_i might be singular which prevents computing the inverse. Therefore, we add a matrix with small diagonal elements to the covariance matrix to prevent the singularity.

3. CLUSTER MERGING OF SYNTHETIC DATA

The LBG and WMD algorithms were used to cluster 2-dimensional data with three Gaussian distributed clusters into six clusters. WMD divided the data as shown in Fig. 1(a) and LBG classified the data points as shown in Fig. 2(a). The equi-Mahalanobis distance ellipses are shown in Fig. 1(a); note that they follow the shape of the actual cluster. Table 1 shows the Mutual Mahalanobis distance between the cluster centers where

$$D_{ij}^M = ||v_i - v_j||^2_{A_i} = (v_i - v_j)' A_i^{-1}(v_i - v_j) \quad (5)$$

is the entry on the i^{th} row and the j^{th} column in Table 1. On the other hand, Table 3 lists the relative Euclidean distance between cluster centers. Unlike the Euclidean distance case, the Mahalanobis distance between cluster center i to cluster center j, D_{ij}^M, is not equal to the Mahalanobis distance between cluster center j to cluster center i, D_{ji}^M. That is because the covariance matrix used in the computation of Eqn. (5) is different.

One important question in any clustering algorithm is how many underlying subgroups are present in the data set. Many algorithms start by making assumptions about the number of clusters, which is sometimes difficult due to lack of prior knowledge. Therefore, estimation of the optimal number of substructures in the data set is a crucial point. One way to determine the number of substructures in the data automatically is

Table 1: Mahalanobis Distance between cluster centers generated using WMD algorithm

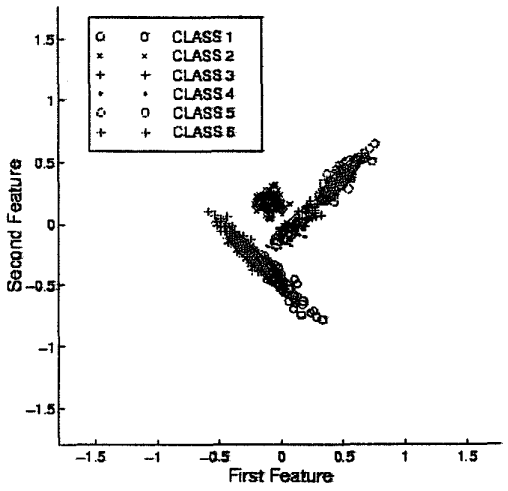
Class	1	2	3	4	5	6
1	0	0.655	0.966	0.370	3.875	0.034
2	0.4003	0	0.077	0.083	0.362	0.184
3	0.3238	0.190	0	0.013	0.081	0.093
4	0.2024	0.158	0.021	0	0.266	0.060
5	0.4186	0.377	0.051	0.097	0	0.289
6	0.0297	0.784	1.132	0.465	4.262	0

Table 2: The ratio of the Mahalanobis Distances between cluster centers generated using WMD algorithm

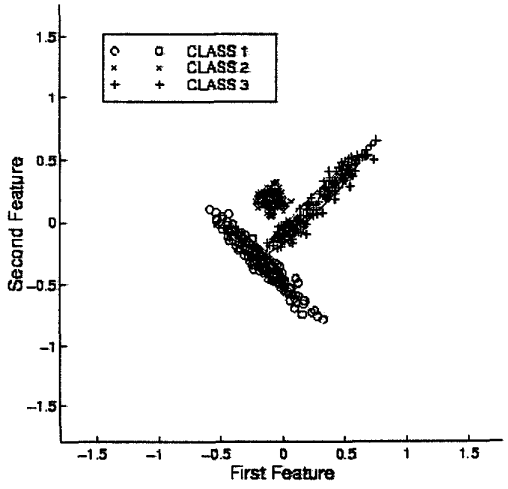
Class	1	2	3	4	5	6
1	0	0.6105	0.3351	0.5467	0.1080	0.8546
2	0	0	0.4076	0.5271	0.9611	0.2348
3	0	0	0	0.6138	0.6294	0.0826
4	0	0	0	0	0.3676	0.1300
5	0	0	0	0	0	0.0680
6	0	0	0	0	0	0

to start with a large number of clusters and have the algorithm merge clusters which meet some criteria of similarity and closeness. Hence the algorithm merges clusters that are believed to be representing the same class and the final number of clusters is the number of distinct classes in the data.

The proposed algorithm merges clusters based on the mutual Mahalanobis distance between cluster centers. We can see from table 1 that class 1 represented by solid circles and class 6 represented by dotted plus signs have small Mahalanobis distance between their cluster centers. This means that members of class 6 can be easily added to class 1 and vice versa. If we look at the Mahalanobis distance between class 6 and class 5 we can see that cluster center 6 is close to cluster center 5 but the opposite is not true. In other words, D_{56}^M is very small but D_{65}^M is large. This can be expected since cluster 5 shows a spread of data points in the direction of cluster center 6 while cluster 6 variance in the direction of cluster center 5 is very small. Therefore, a condition based on the ratio of the Mahalanobis distance between cluster centers was deemed necessary to make sure that the two clusters are not only close

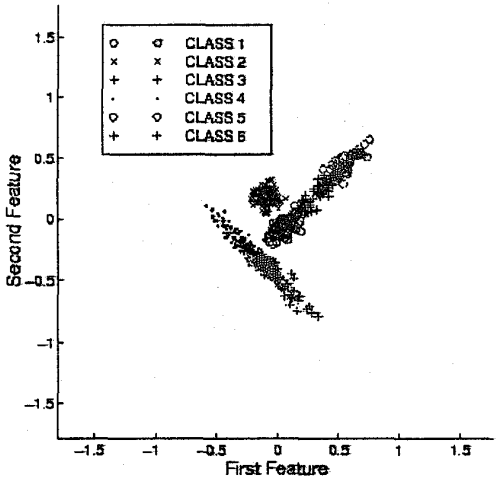


(a)

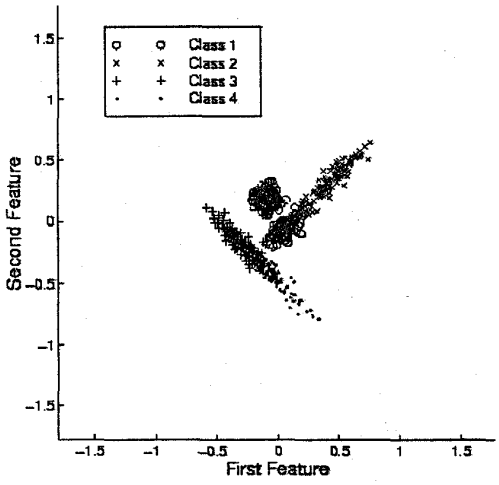


(b)

Figure 1: Results of clustering the data set using WMD algorithm (a) before merging (b) after merging



(a)



(b)

Figure 2: Results of clustering the data set using LBG algorithm (a) before merging (b) after merging

Table 3: Euclidean Distance between cluster centers generated using LBG algorithm

Class	1	2	3	4	5	6
1	0	0.275	0.388	0.381	0.688	0.405
2	0	0	0.428	0.419	0.685	0.625
3	0	0	0	0.752	0.300	0.765
4	0	0	0	0	1.047	0.362
5	0	0	0	0	0	1.058
6	0	0	0	0	0	0

to one another but also similar in orientation.

Table 3 shows the ratio of the Mahalanobis distances between the different cluster centers. Therefore, in order to merge the i^{th} and j^{th} clusters both D_{ij}^{M} and D_{ji}^{M} need to be below a threshold, 0.1 in this case, and also the ratio between D_{ij}^{M} and D_{ji}^{M} needs to be larger than a certain parameter which is selected to be 0.6. Note that the selected parameters are not data dependent since any data set to be clustered are normalized first. Normalization makes the mean of the data set equals zero and the maximum length of data vectors a unity which maintains the shape and relative distance in the data set. However, some control over merging can be achieved by relaxing the choice of these two parameters. Figure 1(b) shows the result of WMD clustering after the clusters have been merged. Note that during cluster merging phase it was decided to merge clusters 3 and 4 together and clusters 3 and 5 also. Therefore, all three clusters were merged together.

In the case of LBG we can see that the only piece of information available for determining the relationship between two clusters is Euclidean distance between their cluster centers. On the other hand, in the case of WMD both of the Mahalanobis distances from each cluster center to the other and also the ratio of these distances can be utilized to infer information about the shape of the clusters. Fig. 2(b) shows the result of LBG clustering after the clusters have been merged. Figure 2 demonstrates the tendency of Euclidean distance clustering to split elongated clusters, classes 3 and 4 in Fig. 2(b), and to merge clusters if the cluster centers are close to each other without regards to their shape, classes 1 and 2 in Fig. 2(a).

We can see that WMD outperforms LBG and would similarly outperform any clustering technique that uses Euclidean distance. Furthermore, Mahalanobis distance degenerates to Euclidean distance for hyperspherical clusters as shown by the circular cluster in Fig. 2(b).

4. BREAST CANCER MAMMOGRAPHIC MASS EXTRACTION

We also applied WMD to the important field of medical imaging. Specificly, we automatically extract suspicious mass-like densities in a mammogram retaining the shape information for the purpose of classification these densities as malignant or benign tumors. Important features that can help classify malignant from benign densities can be derived given that the mass boundaries can be determined [7][8]. Fig. 3 shows a part of digitized mammogram that was passed through a Gaussian filter to smooth the image. Therefore, a thresholding procedure was implemented to extract the bright pixels and the coordinates of these pixels were clustered using both WMD and LBG algorithms. To allow for detecting different size tumors, four clusters were generated and cluster merging was applied to determine the number of actual densities. As we can see from Fig 3, WMD algorithm was able to merge the four clusters given that they were close in shape and the Mahalanobis distances between cluster centers were small. On the other hand, LBG didn't group any of the clusters because of the large size of this tumor, see Fig. 4.

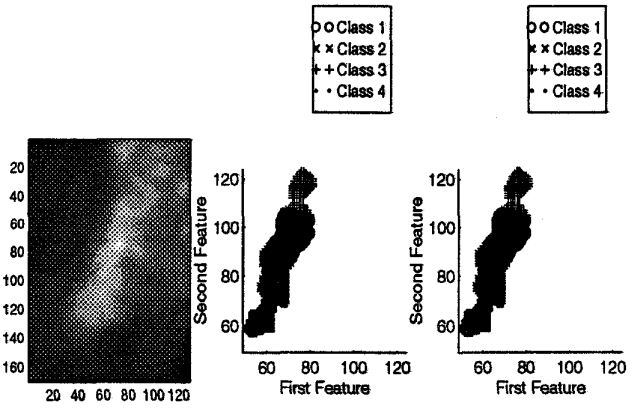


Figure 4: A cancerous tumor and the results of clustering bright pixels using LBG algorithm with 4 clusters before merging and after merging.

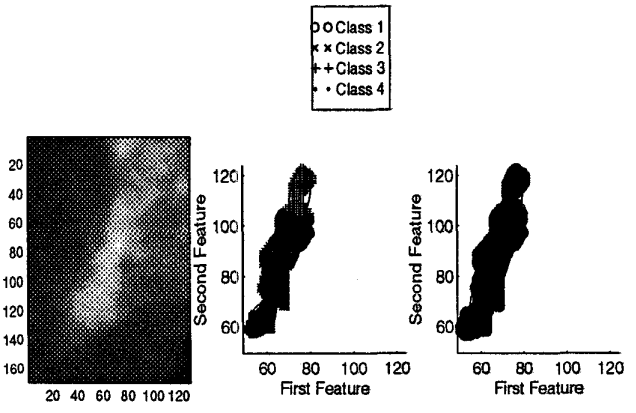


Figure 3: A cancerous tumor and the results of clustering bright pixels using WMD algorithm with 4 clusters before merging and after merging.

To show the effect of clustering different size tumors, another mammogram was tried. In this case, a small cancerous tumor is shown in the left side of Fig. 5 and a benign density is located to the right of it.

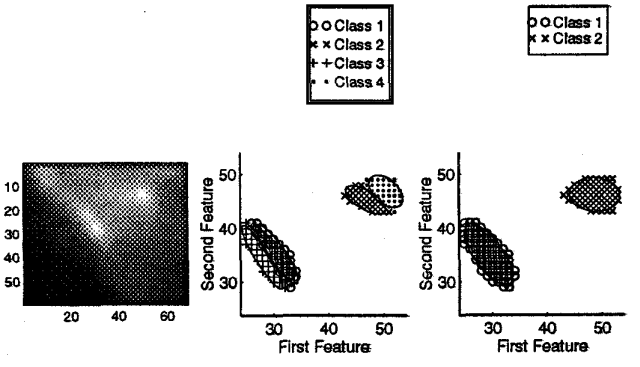


Figure 5: A cancerous tumor and the results of clustering bright pixels using WMD algorithm with 4 clusters before merging and after merging.

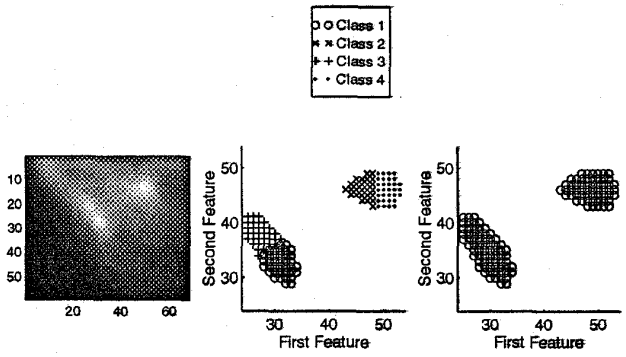


Figure 6: A cancerous tumor and the results of clustering bright pixels using LBG algorithm with 4 clusters before merging and after merging.

Once again, the four ellipsoidal clusters generated by WMD were correctly reduced to two distinct classes. Note that important information about the two densities like cluster center, orientation, variation and size are kept for feature extraction. On the other hand, the LBG algorithm merged the four clusters as one class with the cluster center located in the center, see Fig. 6.

5. CONCLUSION

Mahalanobis Distance is a very useful tool which is believed to give a better measure of similarity than the Euclidean distance. WMD is a simple algorithm useful with hyperellipsoidal data sets as demonstrated by synthetic and real examples. We have shown that the mutual Mahalanobis distances between the cluster centers m_i and m_j using both A_i and A_j are good indicators of the similarity between the two clusters in terms of shape and orientation. These indicators are the basis for merging similar clusters which results in "natural" clustering with automatic determination of the optimal number of clusters. Thus, WMD provides an excellent choice for unsupervised clustering applications.

6. REFERENCES

[1] Allen Gersho and Robert M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[2] J. C. Bezdek, "Self-organization and clustering algorithms," tech. rep., Defence Technical Information Center (DTIC) N91-21783.

[3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* John Wiley and Sons, 1973.

[4] Younis, Khaled S. *Weighted Mahalanobis distance For Hyper-ellipsoidal Clustering.* Master's Thesis, The Air Force Institute of Technology, Dayton, OH, 1996.

[5] J. Mao and A. K. Jain, "A self-organizing network for hyperellipsoidal clustering," *IEEE Transactions on Neural Networks*, vol. 7, pp. 16–29, Jan. 1996.

[6] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, 1980.

[7] G. Svane, E. J. Potchen, A. Sierra, and E. Azavdo, *Screening Mammography: Breast Cancer Diagnosis in Asymptomatic Women*, Mosby-Year Book, Missouri, 1993.

[8] W. A. Polakowski, D. Cournoyer,S. Rogers, M. DeSimio, D. Ruck, J. Hoffmeister, and R. Raines,"Computer-Aided Breast Cancer Detection and Diagnosis of Masses Using Difference of Gaussians and Derivative-Based Feature Saliency", *IEEE Transactions on Medical Imaging*, vol. 16, no. 6, pp. 81–819, 1997.

7. AUTHOR BIOGRAPHIES

Captain Khaled Younis is with the Royal Jordanian Air Force. He received the BS degree in Electrical Engineering (Top Graduate) from Mu'tah University, Jordan, in 1990 and received the MS in Electrical Engineering (Distinguished Graduate) from The Air Force Institute of Technology, Ohio, in 1996. He received a scholarship from the Dayton Area Graduate Studies Institute (DAGSI) to pursue a PhD degree in Electrical Engineering at the University of Dayton with emphasis on pattern recognition. His research interests are in automatic target recognition of images using neural networks, image processing, and signal analysis. He is a member of IEEE and the honor societies Eta Kappa Nu and Tau Beta Pi.